1

Flow Graphs, their Fusion and Data Analysis

Zdzisław Pawlak

Institute of Computer Sciences
Warsaw University of Technology
Ul. Nowowiejska 15/19, 00665 Warsaw, Poland
and
Warsaw School of Information Technology
ul. Newelska 6, 01-447 Warsaw, Poland
zpw@ii.pw.edu.pl

Summary. This paper concerns a new approach to data analysis based on information flow distribution study in flow graphs. The introduced flow graphs differ from that proposed by Ford and Fulkerson, for they do not describe material flow in the flow graph but information "flow" about the data structure.

Data analysis (mining) can be reduced to information flow analysis and the relationship between data can be boiled down to information flow distribution in a flow network. Moreover, it is revealed that information flow satisfies Bayes' rule, which is in fact an information flow conservation equation. Hence information flow has probabilistic character, however Bayes' rule in our case can be interpreted in an entirely deterministic way, without referring to *prior* and *posterior* probabilities, inherently associated with Bayesian philosophy.

Furthermore in this paper we study hierarchical structure of flow networks by allowing to substitute a subgraph determined by branches x and y by a single branch connecting x and y, called *fusion* of x and y. This "fusion" operation allows us to look at data with different accuracy and move from details to general picture of data structure.

Key words: flow graphs, data fusion, data mining, Bayes' rule

1.1 Introduction

In [4] we presented a new approach to data analysis based on information flow distribution study in flow graphs. The introduced flow graphs differ from that proposed by Ford and Fulkerson [1], for they do not describe material flow in the flow graph but information "flow" about the data structure.

With every branch of the flow graph three coefficients are associated, called *strength*, *certainty* and *coverage* factors. These coefficients were widely used in data mining and rough set theory, but in fact they were first introduced by Łukasiewicz [2] in connection with his study of logic and probability. These coefficients have a

probabilistic flavor, but here they are interpreted in a deterministic way, describing information flow distribution in the flow graph.

We claim that data analysis (mining) can be reduced to information flow analysis and the relationship between data can be boiled down to information flow distribution in a flow network. Moreover, it is revealed that information flow satisfies Bayes' rule, which is in fact an information flow conservation equation. Hence information flow has probabilistic character, however Bayes' rule in our case can be interpreted in an entirely deterministic way, without referring to *prior* and *posterior* probabilities, inherently associated with Bayesian philosophy.

Furthermore in this paper we study hierarchical structure of flow networks by allowing to substitute a subgraph determined by branches x and y by a single branch connecting x and y, called *fusion* of x and y. This "fusion" operation allows us to look at data with different accuracy and move from details to general picture of data structure.

This approach allows us to study different relationships in data and can be used as a new mathematical tool for data mining.

Summing up, we advocate to use flow analysis to:

- searching for patterns in data,
- searching for dependencies in data,
- data classification,
- data fusion.

A simple tutorial example will be used to illustrate the introduced ideas.

1.2 Example 1 - Smoking and Cancer

First let us explain basic concepts of the proposed methodology on a simple example taken from [3].

In Table 1.1 data concerning 60 people who do or do not smoke and do or do not have cancer are shown.

Table 1.1. Smoking and Cancer

	Not smoke	Smoke	Total
Not cancer	40	10	50
Cancer	7	3	10
Total	47	13	60

With every data table like that in presented in Table 1.1 we associate a flow graph as shown in Fig.1.1.

Nodes x_0 and x_1 are *inputs* of the graph, whereas y_0 and y_1 are *outputs* of the graph. The numbers assigned to the input nodes $\phi(x_0)$ and $\phi(x_1)$ of the flow graph represent *inflow* to the flow graph, whereas numbers associated with the inputs $\phi(y_0)$ and $\phi(y_1)$ represent *outflow* of the graph. Every branch (x, y) of the flow graph is

labeled by a number which represents a *throughflow* $\phi(x, y)$ through the branch from nodes x to y.

This representation of data is intended to capture the relationships in the data and is not meant to describe any material flow in the network.

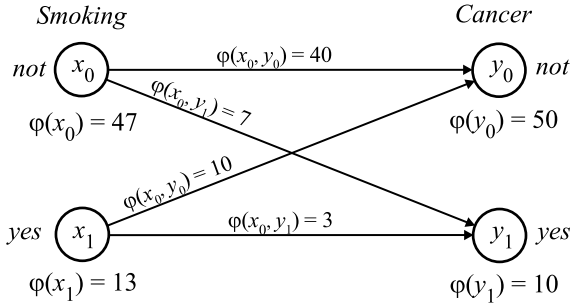

Fig. 1.1. Flow graph for Table 1.1

We will show in the next sections that representation of data as flow in a flow graph can be used to discover many important relationships in data, e.g. dependences. However to this end we have to "normalize" the flow graph by using instead of absolute values of flow $\phi(x)$ their relative values $\sigma(x)$, i.e. percentage of flow with respect to total flow of the graph. The absolute throughflow $\phi(x, y)$ will be also replaced by relative throghflow $\sigma(x, y)$. This normalized representation has very interesting mathematical properties, which can be use to discover patterns in data.

Beside, we will use two additional coefficients called the *certainty* and *coverage factors*, denoted $cer(x, y)$ and $cov(x, y)$ respectively, which characterize how the flow is spread between nodes x and y.

Normalized flow graph for the flow graph given in Fig.1.1 is shown in Fig.1.2.

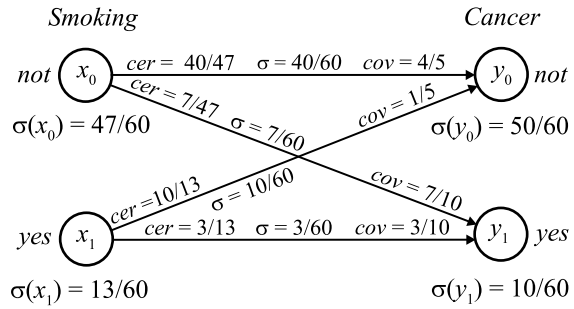

Fig. 1.2. Normalized flow graph for Table 1.1

From the flow graph we arrive at the following conclusions:

- 85% non-smoking persons do not have cancer $(cer(x_0, y_0) = 40/47 \approx 0.85)$,
- 15% non-smoking persons do have cancer $(cer(x_0, y_1) = 7/47 \approx 0.15)$,
- 77% smoking persons do not have cancer $(cer(x_1, y_0) = 10/13 \approx 0.77)$,

- 23% smoking persons do have cancer $(cer(x_1, y_1) = 3/13 \approx 0.23)$.

From the flow graph we get the following reason for having or not cancer:

- 80% persons having not cancer do not smoke $(cov(x_0, y_0) = 4/5 = 0.80)$,
- 20% persons having not cancer do smoke $(cov(x_1, y_0) = 1/5 = 0.20)$,
- 70% persons having cancer do not smoke $(cov(x_0, y_1) = 7/10 = 0.70)$,
- 30% persons having cancer do smoke $(cov(x_1, y_1) = 3/10 = 0.30)$.

Let us observe that in the statistical terminology $\sigma(x_0)$, $\sigma(x_1)$ are *priors* while $\sigma(x_0, y_0), \ldots, \sigma(x_1, y_1)$ are joint distributions, $cov(x_0, y_0), \ldots, cov(x_1, y_1)$ are *posteriors* and $\sigma(y_0), \sigma(y_1)$ are marginal probabilities.

1.3 Flow Graphs Basic Concepts

1.3.1 Flow Graphs

In this section the fundamental concept of the proposed approach flow graph is defined and discussed.

A flow graph is a *directed, acyclic, finite* graph $G = (N, \mathcal{B}, \phi)$, where N is a set of *nodes*, $\mathcal{B} \subseteq N \times N$ is a set of *directed branches*, $\phi : \mathcal{B} \to R^+$ is a *flow function* and R^+ is the set of non-negative reals.

Input of a node $x \in N$ *is the set* $I(x) = \{y \in N : (y, x) \in \mathcal{B}\}$; *output* of a node $x \in N$ is defined as $O(x) = \{y \in N : (x, y) \in \mathcal{B}\}$.

We will also need the concept of *input* and *output* of a graph G, defined, respectively, as follows: $I(G) = \{x \in N : I(x) = \emptyset\}, O(G) = \{x \in N : O(x) = \emptyset\}$.

Inputs and outputs of G are *external nodes* of G; other nodes are *internal nodes* of G.

If $(x, y) \in \mathcal{B}$ then $\phi(x, y)$ is a *throughflow* from x to y.

With every node x of a flow graph G we associate its *inflow*

$$\phi_+(x) = \sum_{y \in I(x)} \phi(y, x), \tag{1.1}$$

and *outflow*

$$\phi_-(x) = \sum_{y \in O(x)} \phi(x, y). \tag{1.2}$$

Similarly, we define an inflow and an outflow for the whole flow graph, which are defined as

$$\phi_+(G) = \sum_{y \in I(G)} \phi_-(x), \tag{1.3}$$

$$\phi_-(G) = \sum_{x \in I(O)} \phi_+(x). \tag{1.4}$$

We assume that for any internal node x, $\phi_+(x) = \phi_-(x) = \phi(x)$, where $\phi(x)$ is a *throughflow* of node x.

Obviously, $\phi_+(G) = \phi_-(G) = \phi(G)$, where $\phi(G)$ is a *troughflow* of graph G.

The above formulas can be considered as *flow conservation equations* [4].

We will define now a *normalized flow graph.*

A normalized flow graph is a *directed, acyclic, finite* graph $G = (N, \mathcal{B}, \sigma)$, where N is a set of *nodes*, $\mathcal{B} \subseteq N \times N$ is a set of *directed branches* and $\sigma : \mathcal{B} \rightarrow < 0, 1 >$ is a *normalized flow* of (x, y) and

$$\sigma(x, y) = \frac{\phi(x, y)}{\phi(G)}, \tag{1.5}$$

is a *strength* of (x, y). Obviously, $0 \leq \sigma(x, y) \leq 1$. The strength of the branch expresses simply the percentage of a total flow through the branch.

In what follows we will use normalized flow graphs only, therefore by flow graphs we will understand normalized flow graphs, unless stated otherwise.

With every node x of a flow graph G we associate its *inflow* and *outflow* defined as

$$\sigma_+(x) = \frac{\phi_+(x)}{\phi(G)} = \sum_{y \in I(x)} \sigma(y, x), \tag{1.6}$$

$$\sigma_-(x) = \frac{\phi_-(x)}{\phi(G)} = \sum_{y \in O(x)} \sigma(x, y). \tag{1.7}$$

Obviously for any internal node x, we have $\sigma_+(x) = \sigma_-(x) = \sigma(x)$, where $\sigma(x)$ is a *normalized throughflow* of x.

Moreover, let

$$\sigma_+(G) = \frac{\phi_+(G)}{\phi(G)} = \sum_{x \in I(G)} \sigma_-(x), \tag{1.8}$$

$$\sigma_-(G) = \frac{\phi_-(G)}{\phi(G)} = \sum_{x \in O(G)} \sigma_+(x). \tag{1.9}$$

Obviously, $\sigma_+(G) = \sigma_-(G) = \sigma(G) = 1$.

If we invert direction of all branches in G, then the resulting graph $G = (N, \mathcal{B}', \sigma')$ will be called an *inverted* graph of G. Of course the inverted graph G' is also a flow graph and all inputs and outputs of G become inputs and outputs of G', respectively.

1.3.2 Certainty and Coverage Factors

With every branch (x, y) of a flow graph G we associate the *certainty* and the *coverage factors*.

The *certainty* and the *coverage* of (x, y) are defined as

$$cer(x, y) = \frac{\sigma(x, y)}{\sigma(x)}, \tag{1.10}$$

and

$$cov(x,y) = \frac{\sigma(x,y)}{\sigma(y)}. \tag{1.11}$$

respectively.

Evidently, $cer(x,y) = cov(y,x)$, where $(x,y) \in \mathcal{B}$ and $(y,x) \in \mathcal{B}'$.

Below some properties, which are immediate consequences of definitions given above are presented:

$$\sum_{y \in O(x)} cer(x,y) = 1, \tag{1.12}$$

$$\sum_{x \in I(y)} cov(x,y) = 1, \tag{1.13}$$

$$\sigma(x) = \sum_{y \in O(x)} cer(x,y)\sigma(x) = \sum_{y \in O(x)} \sigma(x,y), \tag{1.14}$$

$$\sigma(y) = \sum_{x \in I(y)} cov(x,y)\sigma(y) = \sum_{xy \in I(y)} \sigma(x,y), \tag{1.15}$$

$$cer(x,y) = \frac{cov(x,y)\sigma(y)}{\sigma(x)}, \tag{1.16}$$

$$cov(x,y) = \frac{cer(x,y)\sigma(x)}{\sigma(y)}. \tag{1.17}$$

Obviously the above properties have a probabilistic flavor, e.g., equations (14) and (15) have a form of total probability theorem, whereas formulas (16) and (17) are Bayes' rules. However, these properties in our approach are interpreted in a deterministic way and they describe flow distribution among branches in the network.

1.3.3 Paths, Connections and Fusion

A (*directed*) path from x to $y, x \neq y$ in G is a sequence of nodes $x_1, \ldots, x_n$ such that $x_1 = x, x_n = y$ and $(x_i, x_{i+1}) \in \mathcal{B}$ for every $i, 1 \leq i \leq n-1$. A path from x to y is denoted by $[x \ldots y]$.

The *certainty* of the path $[x_1 \ldots x_n]$ is defined as

$$cer[x_1 \ldots x_n] = \prod_{i=1}^{n-1} cer(x_i, x_{i+1}), \tag{1.18}$$

the *coverage* of the path $[x_1 \ldots x_n]$ is

$$cov[x_1 \ldots x_n] = \prod_{i=1}^{n-1} cov(x_i, x_{i+1}), \tag{1.19}$$

and the *strength* of the path $[x \ldots y]$ is

$$\sigma[x \dots y] = \sigma(x)cer[x \dots y] = \sigma(y)cov[x \dots y]. \tag{1.20}$$

The set of all paths from x to $y(x \neq y)$ in G denoted $< x, y >$, will be called a *connection* from x to y in G. In other words, connection $< x, y >$ is a sub-graph of G determined by nodes x and y.

The *certainty* of the connection $< x, y >$ is

$$cer < x, y >= \sum_{[x \dots y] \in < x, y >} cer[x \dots y], \tag{1.21}$$

the *coverage* of the connection $< x, y >$ is

$$cov < x, y >= \sum_{[x \dots y] \in < x, y >} cov[x \dots y], \tag{1.22}$$

and the *strength* of the connection $< x, y >$ is

$$\begin{aligned} \sigma < x, y > &= \sum_{[x \dots y] \in < x, y >} \sigma[x \dots y] = \\ &= \sigma(x)cer < x, y >= \sigma(y)cov < x, y > . \end{aligned} \tag{1.23}$$

If we substitute simultaneously every sub-graph $< x, y >$ of a given flow graph G, where x is an input node and y an output node of G, by a single branch (x, y) such that $\sigma(x, y) = \sigma < x, y >$, then in the resulting graph G', called the *fusion* of G, we have $cer(x, y) = cer < x, y >$, $cov(x, y) = cov < x, y >$ and $\sigma(G) = \sigma(G')$.

Thus fusion of a flow graph can be understood as a simplification of the graph and can be used to get a general picture of relationships in the flow graph.

1.3.4 Dependences in Flow Graphs

Let x and y be nodes in a flow graph $G = (N, \mathcal{B}, \sigma)$, such that $(x, y) \in \mathcal{B}$.

Nodes x and y are *independent* in G if

$$\sigma(x, y) = \sigma(x)\sigma(y). \tag{1.24}$$

From (21) we get

$$\frac{\sigma(x, y)}{\sigma(x)} = cer(x, y) = \sigma(y), \tag{1.25}$$

and

$$\frac{\sigma(x, y)}{\sigma(y)} = cov(x, y) = \sigma(x). \tag{1.26}$$

If

$$cer(x, y) > \sigma(y), \tag{1.27}$$

or

$$cov(x, y) > \sigma(x), \tag{1.28}$$

then x and y are *positively depends* on x in G.

Similarly, if

$$cer(x, y) < \sigma(y), \tag{1.29}$$

or

$$cov(x, y) < \sigma(x), \tag{1.30}$$

then x and y are *negatively dependent* in G.

Relations of independency and dependences are symmetric ones, and are analogous to those used in statistics.

For every branch $(x, y) \in \mathcal{B}$ we define a *dependency* (*correlation*) factor $\eta(x, y)$ defined as

$$\eta(x, y) = \frac{cer(x, y) - \sigma(y)}{cer(x, y) + \sigma(y)} = \frac{cov(x, y) - \sigma(x)}{cov(x, y) + \sigma(x)}. \tag{1.31}$$

Obviously $-1 \leq \eta(x, y) \leq 1$; $\eta(x, y) = 0$ if and only if $cer(x, y) = \sigma(y)$ and $cov(x, y) = \sigma(x)$; $\eta(x, y) = -1$ if and only if $cer(x, y) = cov(x, y) = 0$; $\eta(x, y) = 1$ if and only if $\sigma(y) = \sigma(x) = 0$.

It is easy to check that if $\eta(x, y) = 0$, then x and y are independent, if $-1 \leq \eta(x, y) < 0$ then x and y are negatively dependent and if $0 < \eta(x, y) \leq 1$ then x and y are positively dependent. Thus the dependency factor expresses a degree of dependency, and can be seen as a counterpart of correlation coefficient used in statistics.

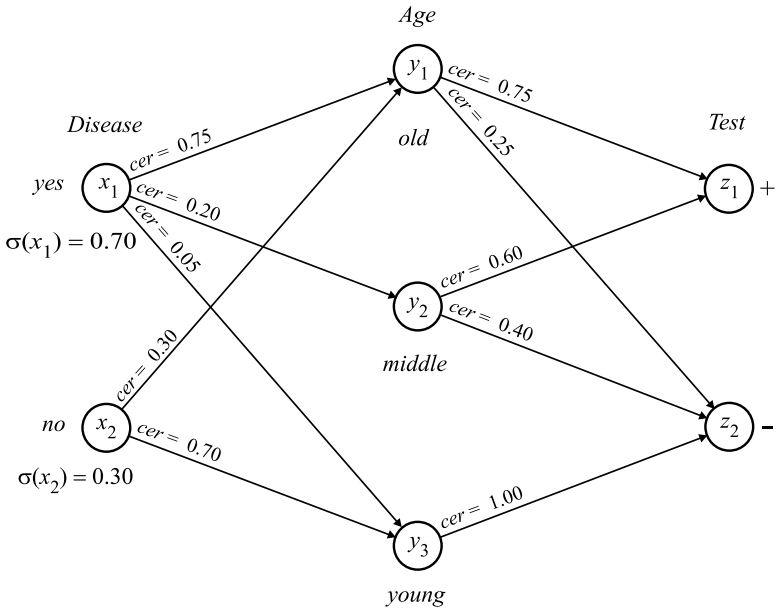

Fig. 1.3. Initial data

1.4 Example 2 - Medical Test

Now we are ready to illustrate the basic concepts presented in this paper by a simple tutorial example.

Various patient groups are put to the test for certain drug effectiveness. Initial data are shown in Fig.1.3. Corresponding flow graph is presented in Fig.1.4.

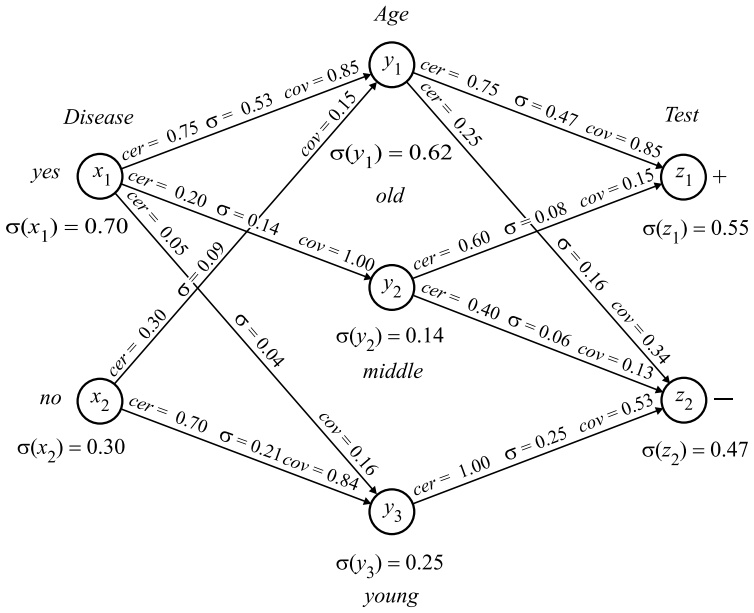

Fig. 1.4. Relationship between *Disease*, *Age* and *Test*

Fig.1.5 shows the corresponding fusion, of *Disease* and *Test*.

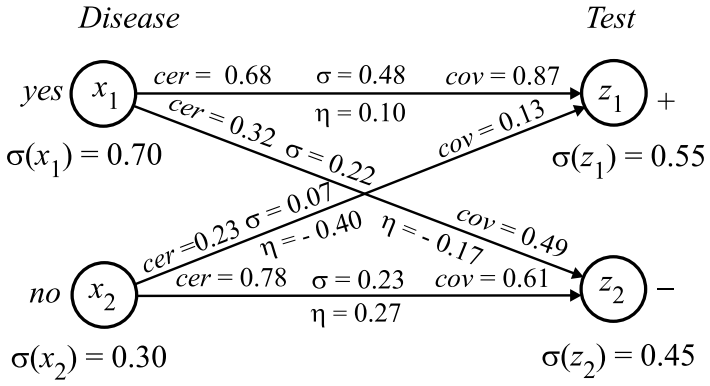

Fig. 1.5. Fusion of the flow graph presented in Fig.1.4

This flow graph leads to the following conclusions:

- If the disease is present then the test result is positive with certainty 0.68
- It the disease is absent then the test result is negative with certainty 0.78

Explanation of test results is as follows:

- If the test result is positive then the disease is present with certainty 0.87
- If the test result is negative then the disease is absent with certainty 0.61

From the flow graph we get:

- There is slight positive correlation between presence of the disease and positive test result $(\eta = 0.10)$.
- There is low positive correlation between absence of the disease and negative test result $(\eta = 0.27)$.
- There is slight negative correlation between presence of the disease and negative test result $(\eta = -0.17)$.
- There is higher negative correlation between absence of the disease and positive test result $(\eta = -0.40)$.

1.5 Conclusions

We proposed in this paper to represent relationships in data by means of flow graphs. Flow in the flow graph is meant to capture structure of data rather than to describe any physical material flow in the network. It is revealed the information flow in the flow graph is governed by Bayes' formula, however the formula can interpreted in entirely deterministic way, without referring to its probabilistic character. This representation allows us to study different relationships in the data and can be used as a new mathematical tool for data mining.

Summing up:

- flow graphs can be used to knowledge representation,
- flow distribution represents relationships in data,
- flow conservation is described by Bayes' formula,
- Bayes' formula has deterministic interpretation.

Acknowledgements

Thanks are due to Professor Andrzej Skowron for critical remarks.

References

1. Ford L.R, Fulkerson D.R,(1962) Flows in Networks. Princeton University Press, Princeton. New Jersey
2. Łukasiewicz J, (1913) Die logishen Grundlagen der Wahrscheinlichkeitsrechnung. Kraków. In: Borkowski L, (ed.), *Jan Łukasiewicz - Selected Works*, North Holland Publishing Company, Amsterdam, London, Polish Scientific Publishers, Warsaw, 1970
3. Grinstead Ch. M, Snell J. L, (1997) Introduction to Probability: Second Revised Edition American Mathematical Society
4. Pawlak Z,(2003) Flow Graphs and Decision Algorithms. In: Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, Proceedings, G. Wang, Y. Yao and A. Skowron (eds.) Lecture Notes in Artificial Intelligence **2639** 1-10 Springer